SCIENCE CHINA
Life Sciences

THEMATIC ISSUE: Computational life sciences
• RESEARCH PAPER •

November 2014 Vol.57 No.11: 1072–1079
doi: 10.1007/s11427-014-4754-7

Incorporating functional annotation information in prioritizing disease associated SNPs from genome wide association studies

HOU Lin[1,2†], MA TianZhou[3†] & ZHAO HongYu[1,2*]

[1]*Department of Biostatistics, Yale School of Public Health, New Haven, CT 06510, USA;*
[2]*VA Cooperative Studies Program Coordinating Center, West Haven, CT 06516, USA;*
[3]*Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA 15261, USA*

Received June 6, 2014; accepted August 12, 2014; published online October 14, 2014

With recent advances in genotyping and sequencing technologies, many disease susceptibility loci have been identified. However, much of the genetic heritability remains unexplained and the replication rate between independent studies is still low. Meanwhile, there have been increasing efforts on functional annotations of the entire human genome, such as the Encyclopedia of DNA Elements (ENCODE) project and other similar projects. It has been shown that incorporating these functional annotations to prioritize genome wide association signals may help identify true association signals. However, to our knowledge, the extent of the improvement when functional annotation data are considered has not been studied in the literature. In this article, we propose a statistical framework to estimate the improvement in replication rate with annotation data, and apply it to Crohn's disease and DNase I hypersensitive sites. The results show that with cell line specific functional annotations, the expected replication rate is improved, but only at modest level.

prioritization, functional annotation, genome wide association studies

Citation: Hou L, Ma TZ, Zhao HY. Incorporating functional annotation information in prioritizing disease associated SNPs from genome wide association studies. Sci China Life Sci, 2014, 57: 1072−1079, doi: 10.1007/s11427-014-4754-7

In recent years, the genetic basis of many traits/diseases have been studied through genome wide association studies (GWAS), resulting in the identifications of tens of thousands of loci that are associated with hundreds of traits/diseases at genome-wide significance level [1]. The number of loci with moderate or suggestive association evidence is even larger, and it is reasonable to anticipate that some of these loci will achieve genome wide significance when more samples are collected and analyzed. It is critical to develop statistical methods to prioritize regions with similar association evidence to improve the replication rate in follow-up studies so as to achieve genome level statistical significance. In this article, we will focus on the analysis of single nucleotide polymorphisms (SNPs) as they constitute the largest class of variants used in GWAS. The procedure to select SNPs for follow-up studies is called post-GWAS prioritization [2], which aims to identify SNPs that are more likely to replicate by incorporating external information. Informative resources include linkage analysis results, GWAS results of a second cohort, biological pathway databases, evidence in medical literature, and functional annotations such as expression quantitative trait loci (eQTL), non-synonymous variation, and others [3].

Intuitively, the external information should prove useful in prioritizing GWAS results and researchers have found evidence that GWAS results are enriched in the sets of SNPs functionally annotated. Minelli et al. [4] conducted a survey among experts in the genetics field, asking them to

†Contributed equally to this work
*Corresponding author (email: hongyu.zhao@yale.edu)

score the importance of different information resources when they choose SNPs for follow-up experiments in their research. The results showed that gene level characteristics, such as "the SNP is in a gene which is highly expressed in a tissue relevant to the phenotype" and "the SNP is in a gene which encodes for a protein in a pathway relevant to the phenotype", were thought as most important by experts. Indeed, prioritization based on gene level evidence has been explicitly employed in GWAS of complex diseases, for example, rheumatoid arthritis [5] and breast cancer [6]. In contrast, functional annotations that are not gene centric, like transcription factor binding sites, DNase I hypersensitive sites, and histone modifications, are not as frequently exploited, despite the considerable biological relevance. In this article, we use functional annotation to refer to a collection of chromosomal regions in the human genome implicated in a functional assay such as ChIP-seq, DHS-seq, and others. There is a need for effective computational approaches to prioritizing GWAS results using functional annotations because 44% of trait/disease susceptibility loci documented in the NHGRI GWAS catalogue as 12/03/13 [1] are located in intergenic regions, which could be overlooked by gene centric methods, and much can be learned of the functional roles of these loci from the rapidly increasing functional annotation data of the non-coding regions. For example, the epigenome mapping by the Encyclopedia of DNA Elements (ENCODE) project [7] and the Roadmap Epigenomics Program [8] has generated large experimental data in a variety of human cell lines and tissues. Researchers hope that these datasets would help to decipher the functional relevance of non-coding SNPs and disease etiology. In addition, prioritization based on functional annotations is more likely to obtain novel findings as opposed to gene-based characteristics, since the latter may be biased to the "known" biology of the trait/disease. There are various web servers that provide annotation information for SNPs on many genomic features [3].

Although proper incorporation of many types of functional annotation data can better prioritize SNPs for follow-up studies, there are a number of unsolved questions that may hinder the developments and applications of effective prioritization strategies using functional annotation data. First, among many types of functional assays in various cell lines, which are more informative for the trait/disease of interest? In other words, how to select appropriate annotation data to prioritize SNPs to increase replication rate in follow-up studies? Second, how much improvement, in terms of replication rate, can we expect by incorporating such information? Obviously, the answers to these questions will depend on the specific diseases to be studied and available functional annotation data. In this paper, we attempt to address these questions by taking Crohn's disease and DNase I hypersensitive sites (DHSs) as an example. We show that when appropriate functional annotation data are incorporated, the replication rate for the prioritized SNPs may be improved. The degree of improvement is annotation specific, but the level of improvement is only modest for the existing annotation data. Therefore, other information and approaches are needed to better prioritize SNPs with a significantly improved replication rate.

1 Methods

1.1 GWAS datasets of Crohn's disease

The *p*-values in the NIDDK study were downloaded from dbGaP (phs000130.v1.p1). The genotype data of the WTCCC cohort were downloaded from the Wellcome Trust Case Control Consortium. In order to match the SNPs with the NIDDK cohort, the genotypes of the WTCCC study were imputed by IMPUTE2 [9], after pre-phasing by SHAPEIT [10]. The Phase 1 haplotypes from the 1000 Genomes project were used as the reference panel [11]. The association *p*-values of the WTCCC study were calculated by linear regression model, after adjusting for population stratification covariates (eq. (1)). In eq. (1), y denotes phenotype, which is 1 for cases and 0 for controls, while x_i is the genotype of SNP i. In the regression model, β_*'s are regression coefficients, and ε is the error term, which follows normal distribution. The first 10 principal components were used to correct for population stratification.

$$y = \beta_0 + \beta x_i + \sum_{s=1}^{10} \beta_s PS_s + \varepsilon. \quad (1)$$

1.2 DNase I hypersensitive sites

DNase I hyper sensitivity is a marker for active *cis*-regulatory elements, including enhancers, promoters, transcription factor binding sites, and other regulatory elements. DNase I hypersensitive sites can be mapped at genome scale by DNase-seq. The analysis results of DNase-seq experiments of three cell lines (GSM1024760, GSM1024790, and GSM736556) were downloaded from the UCSC Genome Browser. The three cell lines are primary Th1, Th17, and normal human epidermal keratinocytes, respectively. Peaks with false discovery rate less than 0.01 were used in the analysis, resulting in 268600, 159605, and 209762 peaks in the whole genome, respectively.

1.3 Kernel estimation of effect size distribution

In order to estimate the effect size distribution of truly associated SNPs in and outside of DHSs, we applied the kernel estimation proposed by Park et al. [12], separately to SNPs in DHSs of Th1 and Th17 cell lines, and to that in the whole genome. Briefly, to obtain an unbiased estimation of the effect size distribution, the effect size of SNPs which are

nominally significant (p-value less than 0.01) in the NIDDK cohort was calculated in the WTCCC cohort. Here, effect size is defined in eq. (2), where $\hat{\beta}$ is the regression coefficient in eq. (1), and f is the corresponding minor allele frequency in the WTCCC cohort.

$$ES = 2\hat{\beta}^2 f(1-f). \tag{2}$$

The density of the effect size distribution was estimated by kernel estimation, using the Gaussian kernel (eq. (3)), where $\phi(\cdot)$ is the density function of standard normal distribution, and h is the bandwidth of the kernel. n is the number of nominally significant SNPs in the NIDDK cohort.

$$\hat{density}(es) = \frac{1}{n}\sum_{i=1}^{n}\phi\left(\frac{es-es_i}{h}\right). \tag{3}$$

1.4 Derivation of expected replication rate (ERR)

The empirical definition of replication rate is the number of replicable SNPs divided by the number of SNPs sent for genotyping in the replication cohort. Accordingly, we define the theoretical replication rate in a probabilistic way, which is the expected number of SNPs that are replicable divided by the number of selected SNPs (eqs. (4) and (5)). We set p_0 to be 0.0001. k is the number of genotyped SNPs in the replication study.

$$ERR = E\left(\sum_{i=1}^{k} I(p_i^{WTCCC} \leqslant p_0)\right)\Big/k, \tag{4}$$

$$\begin{aligned} E(I(p_i^{WTCCC} \leqslant p_0)) &= P(p_i^{WTCCC} \leqslant p_0) \\ &= \int_{es} P(p_i^{WTCCC} \leqslant p_0 \mid es_i = es)\, density(es). \end{aligned} \tag{5}$$

Please note $P(p_i^{WTCCC} \leqslant p_0 \mid es_i = es)$ is the power to detect a SNP with an effect size of es in the WTCCC study. With the effect size given, the power of the study is fixed regardless of whether the SNP is in or outside DHSs. To ease calculation, we divide the range of the effect size distribution into L bins, $\{bin_1, bin_2, \cdots, bin_L\}$, and the value in eq. (5) can be approximated via eq. (6), where the power of the WTCCC study of each bin was estimated as the average power of SNPs that fall into the corresponding interval.

$$\begin{aligned} &E(I(p_i^{WTCCC} \leqslant p_0)) \\ &= \sum_{l=1}^{L} P(es_i \in bin_l)\, Power^{WTCCC}(es_i \in bin_l). \end{aligned} \tag{6}$$

In approach A and B, the effect size distribution was estimated by all SNPs in the genome and SNPs with annotations, respectively (eq. (3)). In approach C, the selected SNPs are a mixture of SNPs with and without annotations, with an unknown mixing proportion m. The expected replication rate is

$$\begin{aligned} ERR &= \sum_{i=1}^{k} P(p_i^{WTCCC} \leqslant p_0)/k \\ &= \sum_{i=1}^{k} P(p_i^{WTCCC} \leqslant p_0, i \in DHS)/k \\ &\quad + \sum_{i=1}^{k} P(p_i^{WTCCC} \leqslant p_0, i \notin DHS)/k \\ &= mP(p_i^{WTCCC} \leqslant p_0, i \in DHS) \\ &\quad + (1-m)P(p_i^{WTCCC} \leqslant p_0, i \notin DHS). \end{aligned} \tag{7}$$

1.5 Calculation of the mixing proportion m with different coverage of annotation

The mixing proportion is calculated by assuming the number of significant SNP with annotation is proportional to the coverage of the annotation. Suppose the fold change of coverage of SNPs with and without coverage is c_1 and c_2, and the mixing proportion in Th1 is m_0, then the mixing proportion is $\dfrac{c_1 m_0}{c_1 m_0 + c_2(1-m_0)}$.

2 Results

2.1 The association signals of Crohn's disease were enriched in DHSs of relevant cell lines

Crohn's disease is a type of inflammatory bowel diseases, which has been intensively studied with GWAS [13–15]. Here we consider the GWAS results from two cohorts, NIDDK and WTCCC (see Materials and methods for details). T-helper cells Th1 and Th17, have been demonstrated to play important roles in the immunopathology of Crohn's disease [16,17], thus it is interesting to ask whether the transcriptional regulatory elements, marked by DNase I hypersensitivity, in Th1 and Th17 can be used to prioritize association signals of Crohn's disease. The DHSs in Th1 and Th17 cells were mapped in the ENCODE project [18,19]. The DHSs dataset of normal human epidermal keratinocytes was included in this study as negative control, since it is presumably not relevant to Crohn's disease.

If DHSs in Th1 and Th17 are informative and helpful to prioritize SNPs with more replicable GWAS signals, an enrichment of association signals should be observed in such regions. In fact, among the 296561 SNPs in the two GWASs, 14083 and 5378 were mapped into the DHSs of the two cell lines, respectively. The number of significant SNPs in the genome and that in the DHSs of Th1 and Th17 are shown in Table 1. We observed a significant enrichment of association signals at different significance cut-offs for both NIDDK and WTCCC studies in DHSs regions compared to the average genome (Figure 1). At a stringent significance level (p-value less than 1×10^{-5}), there were more than two times of association signals in the DHS regions. In contrast, no enrichment was observed in the negative con-

Table 1 A list of number of SNPs in Th1 and Th17 DHSs at different *p*-value cutoffs[a)]

	# SNPs in genome	# SNPs in Th1 DHS	# SNPs in Th17 DHS
All SNPs	296561	14083	5378
SNPs with *p*-value* less than 0.01	4002	221	79
SNPs with *p*-value less than 0.001	540	35	13
SNPs with *p*-value less than 0.0001	92	6	2

a) *, The *p*-values are from the NIDDK GWAS study of Crohn's disease.

trol set. These results underscored the importance of using disease-specific information when incorporating functional annotations.

2.2 The empirical replication rate is higher in DHSs

The enrichment of association signals in a genomic feature simply means that if a SNP is randomly sampled from among the SNPs in the annotated regions, and another SNP is randomly sampled from the entire genome, the former one is more likely to be disease associated. However, a more appropriate scenario in prioritization is given two SNPs with similar *p*-values and different annotations, would the one with annotation be more replicable? We first used empirical data to answer this question. The NIDDK dataset is treated as the discovery panel, and the WTCCC dataset as the replication panel, since the latter has a larger sample size. A SNP is considered "replicable" if its association *p*-value is less than 0.0001 in the WTCCC dataset. As shown in Figure 2A, at fixed association level in the discovery cohort (varying between 0.01 and 0.001), the SNPs in DHSs of Th1 and Th17 were both more replicable, with a fold change around 2. As a consequence, when selecting SNPs to include in the replication panel, we can apply a looser threshold for SNPs in DHSs and a more stringent threshold for SNPs outside DHSs, and obtain comparable replication rates in both sets.

2.3 Theoretical investigation of the improved replication rate in DHSs

Next, we set out to an analytical study for the observation that at a fixed significance level, SNPs in DHSs are more

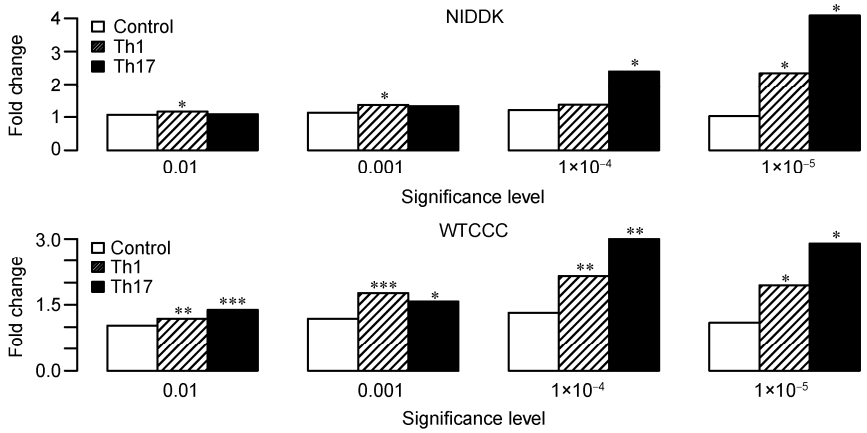

Figure 1 Enrichment of significant SNPs in DHSs of Th1 and Th17 cell lines. *, **, and *** represent significance levels 0.05, 0.01, and 0.001 of the Binomial test, respectively.

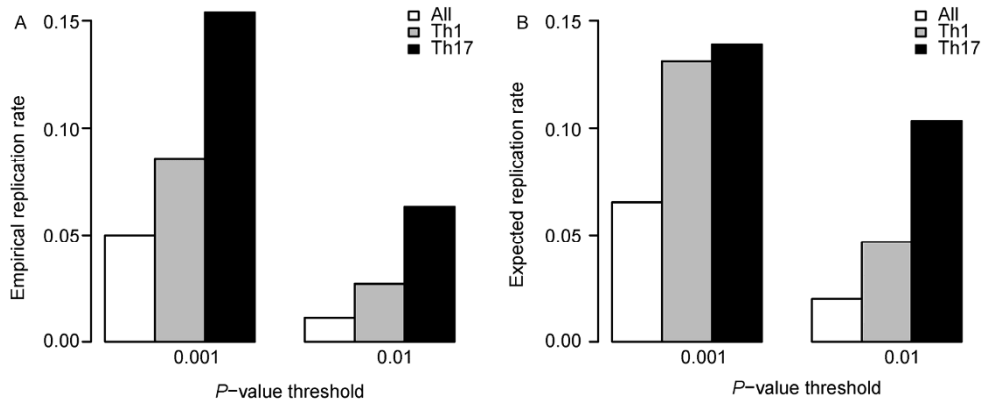

Figure 2 The empirical (A) and expected (B) replication rate of selected SNPs in the genome and that in DHSs of Th1 and Th17 cell lines respectively.

replicable. The necessity for a theoretical framework lies in (i) it leads to a better understanding of the source for the increased replication rate in DHSs; and (ii) it provides estimation of the expected replication rate for a range of prioritization methods.

In a specific ranking approach of the genome-wide SNPs, we took the top k SNPs and considered the expected number of replicable SNPs among them. Here, a SNP is "replicable" if the corresponding p-value in the replication cohort is lower than a given threshold. In the Crohn's disease study, we used 0.0001 as the threshold. The probability that a SNP is replicable depends on two factors: its effect size and the statistical power of the replication study. We adopted the non-parametric estimation of Park et al., to estimate the distribution of the effect sizes of associated SNPs with annotation and that in the whole genome (see Materials and methods). The effect size distributions of SNPs in Th1 and Th17 DHSs are slightly skewed to the right compared to the whole genome (Figure 3). Combined with the power calculation, the expected replication rate in DHSs and the genome can be calculated (see Materials and methods).

In order to investigate whether or not prioritization by DHSs can improve the replication rate, we considered the following ranking approaches: (A) all SNPs that reach certain significance threshold are selected for follow-up studies. Here the original p-values in the NIDDK cohort ($\{p_i, i=1,..., N\}$, N is the number of SNPs in the study) are considered, $p_i^A = p_i$; (B) only SNPs within DHSs and with p-values below the threshold are considered (eq. (8)); (C) the p-values of the SNPs within DHSs are adjusted by a factor of ω (ω>1) while the p-values of the SNPs outside DHSs are kept intact. Then all SNPs of which the adjusted p-values are below the significance threshold are selected (eq. (9)).

$$p_i^B = \begin{cases} 1, & SNP_i \notin DHSs, \\ p_i, & SNP_i \in DHSs. \end{cases} \tag{8}$$

$$p_i^C = \begin{cases} p_i, & SNP_i \notin DHS, \\ p_i / \omega, & SNP_i \in DHS. \end{cases} \tag{9}$$

Approach A represents the scenario of no prioritization, and approaches B and C represent different prioritization methods that incorporate annotation. Apparently, in the approach C, when ω is very large, it reduces to the approach B. In particular, approach C is more practical since we want to loosen the inclusion criteria of SNPs in DHSs in follow-up experiments, but not to exclude SNPs outside DHSs with promising p-values (e.g., SNPs that reach genome-wide significance).

The expected replication rates of the three ranking approaches can be estimated from the effect size distribution and the power in the WTCCC cohort. When the significance threshold in the NIDDK cohort was set to 0.001 (Figure 2B), the expected replication rate of the approach A (no prioritization is performed) was 0.065. In the approach B, the expected replication rates by DHSs in Th1 and Th17 cell lines were 0.131 and 0.139, respectively. Compared to the empirical replication rates (Figure 2), the theoretical results were consistent with the empirical observation that SNPs within DHSs are more replicable. The higher replication rate of SNPs in DHSs should be attributed to an elevated level of effect sizes, since those SNPs have more power to be identified in subsequent studies. In approach C, the selected SNPs are a combination of SNPs within and outside DHSs, and the proportion of mixing varies between 0 and 1, depending on the choice of ω and the p-value distribution in the discovery cohort. Suppose the mixing proportion of SNPs in DHSs is m and the expected replication rates in and outside DHSs are ERR_B and ERR_A, then the expected replication rate of approach C is

$$ERR_C = mERR_B + (1-m)ERR_A. \tag{10}$$

Please note that due to p-value adjustment, the selected SNPs in DHSs have larger p-values compared to those outside DHSs. When we fix the p-value threshold to 0.001, ERR_A is 0.065, while ERR_B varies by the choice of ω (Figure 4A). When ω is between 1 and 4, the expected replication rate in both Th1 and Th17 cell lines is greater than the baseline level (0.065). In other words, with proper choice of ω, approach C can select more SNPs without sacrificing the replication rate. Thus, the DHSs in Th1 and Th17 cell lines are informative in improving the replication rate in Crohn's disease.

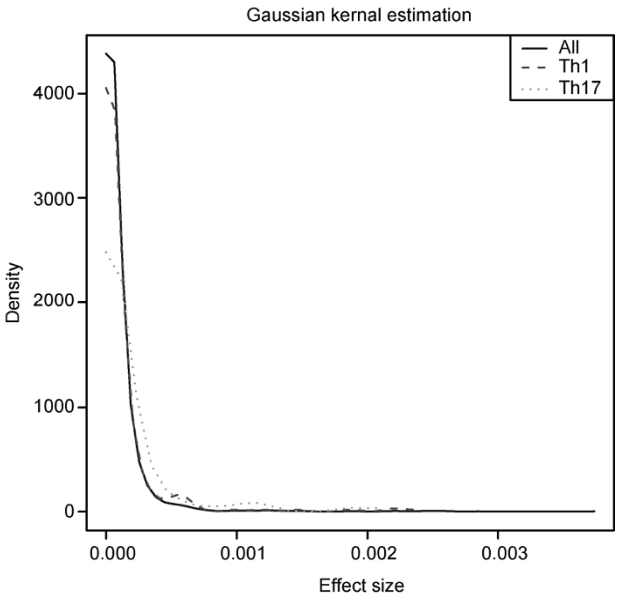

Figure 3 Kernel density estimation of the effect size distribution in the WTCCC cohort. Solid line, nominally significant SNPs in the genome; dashed line, nominally significant SNPs in the DHSs of Th1; dotted line, nominally significant SNPs in the DHSs of Th17.

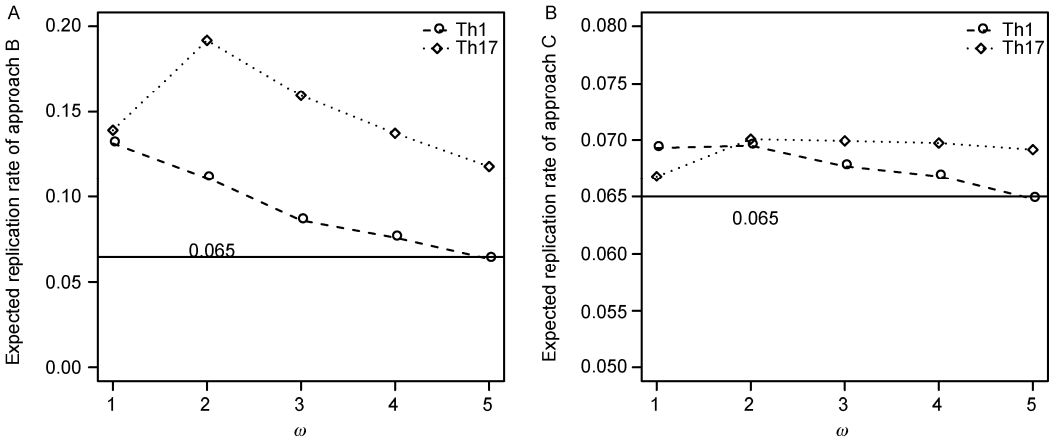

Figure 4 The expected replication rate when varying parameters. A, The expected replication rate of approach B when varying ω. B, The expected replication rate of approach C when varying ω.

Another key question is how much improvement can be expected. In theory, m can take any value between 0 and 1, and a larger m corresponds to a higher replication rate. In the Crohn's disease study, we used empirical data to estimate m. Although the DHSs in Th1 and Th17 cell lines cover only 4.75% and 1.81% of genomic markers, the mixing proportions in these genomic regions are greater due to the enrichment of association signals at these sites. Actually, when ω is set to 4, which is the largest choice of ω with ERR_B greater than the baseline level in the Th1 cell line, m is around 0.16 for the Th1 cell line, and around 0.07 for the Th17 cell line. The expected replication rates were 0.067 and 0.070 for Th1 and Th17 cell lines, respectively (Figure 4B). Thus, the improvement of replication rate was only moderate. A total of 62 and 24 SNPs from DHSs of Th1 and Th17 can be selected, resulting in four and three SNPs that are expected to be replicable respectively.

2.4 Expected replication rate with larger effect size and larger coverage of annotations

In the analysis in Crohn's disease with DHS annotation, the improvement in replication rate was modest, which was limited by (i) the coverage of the annotation and (ii) the extent of the shift of effect size distribution. With the theoretical framework, we can vary these parameters and learn how the expected replication rate changes with them.

We first considered the impact of effect size distribution, which has an effect in the expected replication rate in the annotation. Assume the effect size of an annotation is multiplied by λ as that in Th1 annotation (eq. (11)), the expected replication rate in approach B can be calculated through our theoretical model. The result is 0.164 and 0.261 when λ is set to 2 and 3 correspondingly, and the p-value threshold is 0.001, compared to 0.131 in Th1 DHSs.

$$density^{new}(es) = density^{Th1}(es / \lambda). \quad (11)$$

Next, we assume that the effect size distribution of an annotation is the same as that in Th1 DHSs, but the coverage of the annotation is increased. In this setup, ERR_B is fixed, while the mixing proportion m will increase accordingly. When ω is set to 4, the mixing proportion with Th1 DHSs is 0.16, while its coverage in the genome is 4.75%. When we increase the coverage to 10%, 20%, and 30%, assuming the number of significant SNPs changes proportionally to the coverage, the mixing proportion is increased to 0.30, 0.49, and 0.62 (Methods). As a result (eq. (10)), the overall replication rate will be improved to 0.068, 0.070, and 0.072.

3 Discussion

In this paper, we studied the potential of utilizing functional annotations of the human genome to prioritize GWAS results. The problem was formulated on Crohn's disease and DHSs due to data availability, but the application to other traits and genomic features is straightforward. We found that the choice of disease specific annotation is important for prioritization. The expected replication rate can be improved if the SNPs with annotation have a greater density at regions with large effect size, compared to that in the whole genome. However, the overall improvement may be only moderate, especially when only a limited number of SNPs can be afforded in replication studies. We proposed an analytical framework that calculates the expected replication rate with incorporation of functional annotations. The framework can not only be applied to select functional annotations that can improve replication rate, but also estimate the extent of improvement to be expected.

There are other ways to incorporate biological information other than the two methods we discussed here, which may lead to more optimized result in replication rate. For example, the weights of each marker can vary by the enrichment score of the functional annotation. Nevertheless,

the third ranking method captures the essence of prioritization strategies, which is to put more confidence in SNPs with annotations that are believed to be relevant. There is a great need to develop a most informative prioritization method that can incorporate multiple sources of annotations.

This work was supported in part by the National Institutes of Health (R01 GM59507 and U01 HG005718) and the VA Cooperative Studies Program of the Department of Veterans Affairs, Office of Research and Development. This study makes use of data generated by the Wellcome Trust Case Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the project was provided by Wellcome Trust under award 076113. We also thank dbGaP for the association results of phs000130.v1.p1.

1 Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci USA, 2009, 106: 9362–9367
2 Cantor RM, Lange K, Sinsheimer JS. Prioritizing GWAS results: a review of statistical methods and recommendations for their application. Am J Hum Genet, 2010, 86: 6–22
3 Hou L, Zhao H. A review of post-GWAS prioritization approaches. Front Genet, 2013, 4: 280
4 Minelli C, De Grandi A, Weichenberger CX, Gögele M, Modenese M, Attia J, Barrett JH, Boehnke M, Borsani G, Casari G, Fox CS, Freina T, Hicks AA, Marroni F, Parmigiani G, Pastore A, Pattaro C, Pfeufer A, Ruggeri F, Schwienbacher C, Taliun D, Pramstaller PP, Domingues FS, Thompson JR. Importance of different types of prior knowledge in selecting genome-wide findings for follow-up. Genet Epidemiol, 2013, 37: 205–213
5 Stahl EA, Raychaudhuri S, Remmers EF, Xie G, Eyre S, Thomson BP, Li Y, Kurreeman FAS, Zhernakova A, Hinks A, Guiducci C, Chen R, Alfredsson L, Amos CI, Ardlie KG, Barton A, Bowes J, Brouwer E, Burtt NP, Catanese JJ, Coblyn J, Coenen MJH, Costenbader KH, Criswell LA, Crusius JBA, Cui J, de Bakker PIW, De Jager PL, Ding B, Emery P, Flynn E, Harrison P, Hocking LJ, Huizinga TWJ, Kastner DL, Ke X, Lee AT, Liu X, Martin P, Morgan AW, Padyukov L, Posthumus MD, Radstake TRDJ, Reid DM, Seielstad M, Seldin MF, Shadick NA, Steer S, Tak PP, Thomson W, van der Helm-van Mil AHM, van der Horst-Bruinsma IE, van der Schoot CE, van Riel PLCM, Weinblatt ME, Wilson AG, Wolbink GJ, Wordsworth BP, Wijmenga C, Karlson EW, Toes REM, de Vries N, Begovich AB, Worthington J, Siminovitch KA, Gregersen PK, Klareskog L, Plenge RM. Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. Nat Genet, 2010, 42: 508–514
6 Thomas G, Jacobs KB, Kraft P, Yeager M, Wacholder S, Cox DG, Hankinson SE, Hutchinson A, Wang Z, Yu K, Chatterjee N, Garcia-Closas M, Gonzalez-Bosquet J, Prokunina-Olsson L, Orr N, Willett WC, Colditz GA, Ziegler RG, Berg CD, Buys SS, McCarty CA, Feigelson HS, Calle EE, Thun MJ, Diver R, Prentice R, Jackson R, Kooperberg C, Chlebowski R, Lissowska J, Peplonska B, Brinton LA, Sigurdson A, Doody M, Bhatti P, Alexander BH, Buring J, Lee IM, Vatten LJ, Hveem K, Kumle M, Hayes RB, Tucker M, Gerhard DS, Fraumeni JF, Hoover RN, Chanock SJ, Hunter DJ. A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (RAD51L1). Nat Genet, 2009, 41: 579–584
7 The EPC. A user's guide to the Encyclopedia of DNA Elements (ENCODE). PLoS Biol, 2011, 9: e1001046
8 Chadwick LH. The NIH Roadmap Epigenomics Program data resource. Epigenomics, 2012, 4: 317–324
9 Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet, 2009, 5: e1000529
10 Delaneau O, Howie B, Cox Anthony J, Zagury JF, Marchini J. Haplotype estimation using sequencing reads. Am J Hum Genet, 2013, 93: 687–696
11 The 1000 Genome Project Consortium. An integrated map of genetic variation from 1092 human genomes. Nature, 2012, 491: 56–65
12 Park JH, Wacholder S, Gail MH, Peters U, Jacobs KB, Chanock SJ, Chatterjee N. Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. Nat Genet, 2010, 42: 570–575
13 Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature, 2007, 447: 661–678
14 Franke A, McGovern DPB, Barrett JC, Wang K, Radford-Smith GL, Ahmad T, Lees CW, Balschun T, Lee J, Roberts R, Anderson CA, Bis JC, Bumpstead S, Ellinghaus D, Festen EM, Georges M, Green T, Haritunians T, Jostins L, Latiano A, Mathew CG, Montgomery GW, Prescott NJ, Raychaudhuri S, Rotter JI, Schumm P, Sharma Y, Simms LA, Taylor KD, Whiteman D, Wijmenga C, Baldassano RN, Barclay M, Bayless TM, Brand S, Buning C, Cohen A, Colombel J-F, Cottone M, Stronati L, Denson T, De Vos M, D'Inca R, Dubinsky M, Edwards C, Florin T, Franchimont D, Gearry R, Glas J, Van Gossum A, Guthery SL, Halfvarson J, Verspaget HW, Hugot J-P, Karban A, Laukens D, Lawrance I, Lemann M, Levine A, Libioulle C, Louis E, Mowat C, Newman W, Panes J, Phillips A, Proctor DD, Regueiro M, Russell R, Rutgeerts P, Sanderson J, Sans M, Seibold F, Steinhart AH, Stokkers PCF, Torkvist L, Kullak-Ublick G, Wilson D, Walters T, Targan SR, Brant SR, Rioux JD, D'Amato M, Weersma RK, Kugathasan S, Griffiths AM, Mansfield JC, Vermeire S, Duerr RH, Silverberg MS, Satsangi J, Schreiber S, Cho JH, Annese V, Hakonarson H, Daly MJ, Parkes M. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nat Genet, 2010, 42: 1118–1125
15 Jostins L, Ripke S, Weersma RK, Duerr RH, McGovern DP, Hui KY, Lee JC, Philip Schumm L, Sharma Y, Anderson CA, Essers J, Mitrovic M, Ning K, Cleynen I, Theatre E, Spain SL, Raychaudhuri S, Goyette P, Wei Z, Abraham C, Achkar JP, Ahmad T, Amininejad L, Ananthakrishnan AN, Andersen V, Andrews JM, Baidoo L, Balschun T, Bampton PA, Bitton A, Boucher G, Brand S, Buning C, Cohain A, Cichon S, D'Amato M, De Jong D, Devaney KL, Dubinsky M, Edwards C, Ellinghaus D, Ferguson LR, Franchimont D, Fransen K, Gearry R, Georges M, Gieger C, Glas J, Haritunians T, Hart A, Hawkey C, Hedl M, Hu X, Karlsen TH, Kupcinskas L, Kugathasan S, Latiano A, Laukens D, Lawrance IC, Lees CW, Louis E, Mahy G, Mansfield J, Morgan AR, Mowat C, Newman W, Palmieri O, Ponsioen CY, Potocnik U, Prescott NJ, Regueiro M, Rotter JI, Russell RK, Sanderson JD, Sans M, Satsangi J, Schreiber S, Simms LA, Sventoraityte J, Targan SR, Taylor KD, Tremelling M, Verspaget HW, De Vos M, Wijmenga C, Wilson DC, Winkelmann J, Xavier RJ, Zeissig S, Zhang B, Zhang CK, Zhao H, Silverberg MS, Annese V, Hakonarson H, Brant SR, Radford-Smith G, Mathew CG, Rioux JD, Schadt EE, Daly MJ, Franke A, Parkes M, Vermeire S, Barrett JC, Cho JH. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature, 2012, 491: 119–124
16 Brand S. Crohn's disease: Th1, Th17 or both? The change of a paradigm: new immunological and genetic insights implicate Th17 cells in the pathogenesis of Crohn's disease. Gut, 2009, 58: 1152–1167
17 Strober W, Fuss IJ. Proinflammatory cytokines in the pathogenesis of inflammatory bowel diseases. Gastroenterology, 2011, 140: 1756–1767.e1
18 Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J, Shafer A, Neri F, Lee K, Kutyavin T, Stehling-Sun S, Johnson AK, Canfield TK, Giste E, Diegel M, Bates D, Hansen RS, Neph S, Sabo PJ, Heimfeld S, Raubitschek A, Ziegler S, Cotsapas C, Sotoodehnia N, Glass I, Sunyaev SR, Kaul R, Stamatoyannopoulos JA. Systematic

localization of common disease-associated variation in regulatory DNA. Science, 2012, 337: 1190–1195

19 Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B, Garg K, John S, Sandstrom R, Bates D, Boatman L, Canfield TK, Diegel M, Dunn D, Ebersol AK, Frum T, Giste E, Johnson AK, Johnson EM, Kutyavin T, Lajoie B, Lee B K, Lee K, London D, Lotakis D, Neph S, Neri F, Nguyen ED, Qu H, Reynolds AP, Roach V, Safi A, Sanchez ME, Sanyal A, Shafer A, Simon JM, Song L, Vong S, Weaver M, Yan Y, Zhang Z, Zhang Z, Lenhard B, Tewari M, Dorschner MO, Hansen RS, Navas PA, Stamatoyannopoulos G, Iyer VR, Lieb JD, Sunyaev SR, Akey JM, Sabo PJ, Kaul R, Furey TS, Dekker J, Crawford GE, Stamatoyannopoulos JA. The accessible chromatin landscape of the human genome. Nature, 2012, 489: 75–82